

1 **Statistical Post-Processing of Ensemble Precipitation Forecasts by Fitting**
2 **Censored, Shifted Gamma Distributions**

3 Michael Scheuerer*

4 *University of Colorado, Cooperative Institute for Research in Environmental Sciences at the*
5 *NOAA Earth System Research Laboratory*

6 Thomas M. Hamill

7 *National Oceanic and Atmospheric Administration (NOAA), Earth System Research Laboratory,*
8 *Physical Sciences Division, Boulder, Colorado*

9 * *Corresponding author address:* Dr. Michael Scheuerer,

10 NOAA/ESRL, Physical Sciences Division, 325 Broadway, R/PSD1, Boulder, CO 80305-3337.

11 E-mail: michael.scheuerer@noaa.gov

ABSTRACT

12 A parametric statistical post-processing method is presented that transforms
13 raw (and frequently biased) ensemble forecasts from the Global Ensemble
14 Forecast System (GEFS) into reliable predictive probability distributions for
15 precipitation accumulations. Exploratory analysis based on 12 years of refore-
16 cast data and 1/8-degree climatology-calibrated precipitation analyses shows
17 that censored, shifted gamma distributions can well approximate the condi-
18 tional distribution of observed precipitation accumulations given the ensem-
19 ble forecasts. A nonhomogeneous regression model is set up to link the pa-
20 rameters of this distribution to ensemble statistics that summarize the mean
21 and spread of predicted precipitation amounts within a certain neighborhood
22 of the location of interest, and in addition the predicted mean of precipitable
23 water. The proposed method is demonstrated with precipitation reforecasts
24 over the conterminous United States using common metrics such as Brier
25 skill scores and reliability diagrams. It yields probabilistic forecasts that are
26 reliable, highly skillful, and sharper than the previously demonstrated analog
27 procedure. In situations with limited predictability, increasing the size of the
28 neighborhood within which ensemble forecasts are considered as predictor
29 can further improve forecast skill. It is found, however, that even a parametric
30 post-processing approach crucially relies on the availability of a sufficiently
31 large training dataset.

32 1. Introduction

33 Ensemble predictions are now routinely generated at operational weather prediction centers
34 worldwide (Molteni et al. 1996; Toth and Kalnay 1993, 1997; Houtekamer and Derome 1995;
35 Charron et al. 2010). Despite many improvements to them over the last ~ 2 decades, precipitation
36 forecasts from the ensembles are still typically unreliable, be it from insufficient model resolution,
37 less-than-optimal initial conditions, sub-optimal treatment of model uncertainty, and/or sampling
38 error. For this reason, statistical post-processing of the output of an ensemble prediction system
39 is commonly an integral part of the forecast process, since it can improve the reliability and skill
40 of probabilistic guidance (e.g. Wilks and Hamill 2007; Hamill et al. 2008, and references therein).
41 By comparing past forecasts with their verifying observations, systematic biases and inadequate
42 representation of forecast uncertainty can be identified, and the current forecast can be adjusted
43 such as to minimize these systematic errors. When the forecasts are provided on a grid that is too
44 coarse to resolve small-scale effects that affect the weather variable under consideration, many
45 post-processing methods also implicitly perform a statistical downscaling.

46 The statistical post-processing of precipitation accumulations is far more challenging than the
47 post-processing of weather variables like surface temperature or wind speed for several reasons:

- 48 1. Their mixed discrete/continuous nature (positive probability of being exactly zero, contin-
49 uous value range for positive precipitation amounts) makes it difficult to find an adequate
50 parametric distribution model.
- 51 2. Forecast uncertainty typically increases with the magnitude of expected precipitation
52 amounts; this must be taken into account when setting up a model for the conditional dis-
53 tribution of observed precipitation amounts given the ensemble forecasts.

3. High precipitation amounts occur very infrequently; a customized treatment of these cases may therefore require a vast amount of training data.

The advantages and disadvantages of the different post-processing approaches proposed in the literature are typically related to those three challenges. Non-parametric approaches like the analog method (Hamill and Whitaker 2006; Hamill et al. 2015) completely avoid the first two issues, but may be disproportionately affected by the third one since their treatment of high precipitation amounts neglects the information with training samples with lower precipitation amounts. Parametric methods, on the other hand, can extrapolate the relations found between observations and forecasts of low and moderate magnitudes to higher magnitudes. In doing so, they may reduce the demand for training data, but the quality of the corresponding predictions strongly depends on the adequacy of the parametric assumptions that have to be made. Examples of parametric approaches that have been developed for quantitative precipitation forecasts include Bayesian Model Averaging (BMA, Sloughter et al. 2007), extended logistic regression (ExLR, Wilks 2009; Ben Bouallègue 2013; Messner and Mayr 2014), and ensemble model output statistics (EMOS, Scheuerer 2014). All of them make somewhat ad-hoc assumptions about the parametric form of the predictive distributions: Sloughter’s BMA method models precipitation occurrence/non-occurrence separately and assumes gamma distributions for positive precipitation amounts; ExLR implies the assumption of censored logistic distributions; Scheuerer’s EMOS method assumes censored generalized extreme value distributions. To deal with the issue of heteroscedasticity mentioned above, BMA and ExLR commonly apply power-transformations to both forecasts and observations, with powers chosen such as to make the forecast error terms more homoscedastic. Scheuerer’s EMOS method utilizes two different ensemble statistics that serve as predictors for the scale parameter of the censored GEV distributions.

77 In this paper we will leverage NOAA’s second-generation GEFS reforecast data set (Hamill et al.
78 2013) to systematically develop a parametric model for the conditional distribution of observed
79 precipitation amounts given the ensemble forecasts. This will eventually lead to an approach sim-
80 ilar to the one proposed by Scheuerer (2014), but based on censored, shifted gamma distributions
81 (hereafter, CSGD), and a more sophisticated heteroscedastic regression model that accounts for
82 some further peculiarities of precipitation. In Section 2 we briefly describe the forecast and obser-
83 vation data used in this study, and we introduce our CSGD model in Section 3. Section 4 describes
84 the actual post-processing approach, which proceeds in three steps: first, the ensemble forecasts
85 are adjusted such as to match the observation climatology, and are condensed into three ensemble
86 statistics. Second, a CSGD model for the unconditional (climatological) distribution of the obser-
87 vations is fitted. Finally, a nonhomogeneous regression model is set up which links the ensemble
88 statistics to the CSGD parameters, and results in a conditional distribution model for the obser-
89 vations given the ensemble forecasts. This model is relatively complex, but a comparison with
90 non-parametrically estimated conditional distributions of observed precipitation amounts shows
91 that a certain degree of flexibility (and thus complexity) is necessary to address the peculiarities
92 of precipitation. The benefit of developing a sophisticated parametric approach will become clear
93 in Section 5, where probabilistic forecasts generated by our method are verified and compared
94 against those obtained with a state-of-the-art analog approach. The latter is non-parametric, thus
95 even more flexible than the CSGD approach, and easier to implement. In situations where training
96 data is sparse (e.g. rare events), however, the predictive performance of the CSGD method is fa-
97 vorable. Further experiments are presented that study how the different components of the CSGD
98 contribute to the overall performance, and how reducing the amount of training data affects the
99 quality of the fitted regression model. Section 6 provides a summary and points out challenges
100 with parametric post-processing approaches that require further investigations.

2. Data

The post-processing method developed here is applied to 12-hourly accumulated precipitation forecasts during the period from January 2002 to December 2013 for lead times up to +6 days. All of the forecast data were obtained from the second-generation GEFS reforecast data set; the same data was used in a recent paper by Hamill et al. (2015) which discusses variants of the analog method for statistical post-processing of ensemble precipitation forecasts. For precipitation, individual forecasts by the 11-member GEFS reforecast ensemble were retrieved, and forecast data was extracted on GEFS's native Gaussian grid at $\sim 1/2$ -degree resolution in an area surrounding the contiguous U.S. Total-column ensemble-mean precipitable water is used as an additional predictor in our regression model, and the corresponding forecasts were interpolated to the same grid before further processing. Again as in Hamill et al. (2015), post-processing and verification is performed against precipitation analyses from the climatology-calibrated precipitation analysis (CCPA) data set of Hou et al. (2014), which were obtained on a $\sim 1/8$ -degree grid inside the contiguous U.S. The downscaling from the $\sim 1/2$ -degree to the $\sim 1/8$ -degree resolution will implicitly be part of the post-processing procedure.

3. The censored, shifted gamma distribution

To set up a parametric post-processing method, a suitable class of probability distributions must be identified. As precipitation occurrence/non-occurrence and amount are modeled jointly, a convenient way to do so is using a continuous distribution that permits negative values, and left-censoring it at zero, i.e., replacing all negative values by zero. The censoring turns the probability for negative values of the uncensored distribution into a probability of observing a value equal to zero, thus ensuring requirement 1 described in Section 1.

Exploratory data analysis reveals another challenging requirement for conditional distributions of precipitation accumulations: when the predictor variable (e.g. the ensemble-mean precipitation forecast) is small, then a strongly right-skewed distribution is called for, but the required skewness becomes smaller and smaller as the predictor variable's magnitude increases. To some extent, this behavior can be addressed by using gamma distributions, which are characterized by a shape parameter k and a scale parameter θ . Those two parameters are related to the mean μ and the standard deviation σ of the gamma distribution via

$$k = \frac{\mu^2}{\sigma^2}, \quad \theta = \frac{\sigma^2}{\mu} \quad (1)$$

(Wilks 2011, Sec. 4.4.3). Since the predictive standard deviation increases more slowly than the predictive mean as the predictor variables increase, the shape parameter k decreases, and as k increases the skewness decreases.

A disadvantage of the gamma distribution is that its value range is non-negative. To make the above censoring idea feasible, we therefore introduce an additional parameter $\delta > 0$. This shifts the cumulative distribution function (CDF) of the gamma distribution somewhat to the left. That is, if F_k denotes the CDF of a gamma distribution with unit scale and shape parameter k , then the CDF $\tilde{F}_{k,\theta,\delta}$ of our censored, shifted gamma distribution (CSGD) model is defined by

$$\tilde{F}_{k,\theta,\delta}(y) = \begin{cases} F_k\left(\frac{y-\delta}{\theta}\right) & \text{for } y \geq 0 \\ 0 & \text{for } y < 0 \end{cases} \quad (2)$$

Using the relations in (1), this distribution can also be parametrized by μ , σ , and δ : μ reflects the expected magnitude of precipitation; σ parametrizes prediction uncertainty; δ reduces the magnitude of precipitation somewhat and controls the probability of zero precipitation. An illustration of the CSGD is given in Fig. 1. Note that σ affects both the continuous part of the distribution and the point mass at zero, which we feel is consistent with its interpretation as an uncertainty

parameter: if the expected amount of precipitation is 1 mm, high forecast uncertainty implies that there is still a certain chance of observing much more precipitation, but also a significant chance of observing no precipitation at all. Increasing σ while keeping μ and δ fixed shifts more mass below the censoring threshold and thus accounts for both implications of increased uncertainty. A two-stage approach which models precipitation occurrence and amount separately offers more flexibility, but does not have a single parameter that can be interpreted as uncertainty in this way.

4. Post-processing method

Having selected a family of probability distributions, we propose a procedure to link the three parameters of this distribution to the ensemble forecasts. This is done in three steps. First, quantile mapping is performed to adjust the ensemble precipitation forecasts such as to match the observation climatology. The adjusted forecasts are then reduced to two statistics that measure mean and spread of predicted precipitation accumulations. A further statistic is calculated that measures the mean precipitable water. In the second step, we fit a CSGD model as in eq. (2) to the observed daily climatology of 12-h precipitation accumulations at each grid point (separately for each month). This CSGD model is the basis for a heteroscedastic regression model that is set up in the third step, and links the ensemble statistics from step one to the CSGD parameters, thus defining a predictive distribution for the observed precipitation accumulations, given the ensemble forecasts. We now consider each step in more detail.

a. Quantile mapping and ensemble statistics

As a first step in our post-processing scheme, we attempt to correct systematic errors in ensemble forecast climatology. For example, the underlying numerical weather prediction may produce too many days with light precipitation and underforecast heavy precipitation events. Alternatively,

these errors can arise due to coarser spatial resolution of the forecast grid compared to the grid on which analyzed precipitation is available. This first step can therefore also be viewed as a preliminary downscaling procedure.

Let s be a location associated with some analysis grid point. Prediction errors of the ensemble forecasts may result from inaccurately predicted magnitudes of a precipitation event as described above, but may also be caused by displacement errors. For example, a front or thunderstorm may have been predicted by the numerical weather prediction (NWP) model, but its position might be shifted away somewhat from its true position. The ensemble size of operational ensemble forecast systems is usually too small to represent this position uncertainty, and we therefore follow Scheuerer (2014) and consider ensemble forecasts at all forecast grid points within a certain neighborhood $N(s)$ of s as potential predictors for the analyzed precipitation amount at s . Forecast f_{xj} of ensemble member j at forecast grid point x is thus used multiple times to calculate ensemble- and spatial means and spreads for all analysis grid point neighborhoods $N(s_1), N(s_2), \dots$ containing x . Each time, the forecasts within $N(s)$ are adjusted such that their climatology matches the respective observation climatology as illustrated in Fig. 2. This is achieved via quantile mapping: for each forecast f_{xj} we determine to which quantile $q_{f,x}(p), p \in [0, 1]$ of the forecast climatology it corresponds, and then map it to the corresponding quantile $q_{o,s}(p)$ of the observation climatology. The quantiles are estimated from the training sample; for the GEFS ensemble considered here, all members are exchangeable, can thus be assumed to have the same forecast climatology, and can be pooled for the purpose of estimating the forecast quantiles. Estimating higher quantiles still comes with substantial sampling variability, and to make our quantile mapping procedure more robust, we therefore resort to a linear approximation of the mapping function for quantiles above the 90% quantile (details of this procedure are given in online appendix A).

188 To use the adjusted ensemble forecasts within a regression framework, they are next condensed
 189 into statistics that summarize the most important information. As discussed above, we propose
 190 that all forecast grid points in $N(s)$ - which we take as a neighborhood around s with radius r -
 191 should be considered in determining these statistics, but we still expect forecasts at grid points
 192 closer to s to be more informative about the precipitation at s . Following Scheuerer (2014), we
 193 therefore weigh the forecast grid points according to their distance to s and let

$$w_{sx} \sim \max \left\{ 1 - \left(\frac{\text{dist}(x, s)}{r} \right)^2, 0 \right\}$$

194 with a constant of proportionality chosen such that the weights sum up to one (see left panel of
 195 Fig. 2 for an illustration of this weighting scheme). Assuming that we have an m -member ensemble
 196 with adjusted precipitation forecasts $\tilde{f}_{x1}, \dots, \tilde{f}_{xm}$ and forecasts $\chi_{x1}, \dots, \chi_{xm}$ of precipitable water,
 197 we consider the following ensemble statistics:

$$\text{POP}_{f,s} := \frac{1}{m} \sum_{j=1}^m \sum_{x \in N(s)} w_{sx} \mathbf{1}_{\{\tilde{f}_{xj} > 0\}} \quad (3)$$

$$\bar{f}_s := \frac{1}{m} \sum_{j=1}^m \sum_{x \in N(s)} w_{sx} \tilde{f}_{xj} \quad (4)$$

$$\bar{\chi}_s := \frac{1}{m} \sum_{k=1}^m \sum_{x \in N(s)} w_{sx} \chi_{xk} \quad (5)$$

$$\text{MD}_{f,s} := \frac{1}{m^2} \sum_{j,j'=1}^m \sum_{x,x' \in N(s)} w_{sx} w_{sx'} |\tilde{f}_{xj} - \tilde{f}_{x'j'}| \quad (6)$$

198 The first statistic describes the probability of precipitation derived from the (augmented and
 199 weighted) ensemble. The second and third are the weighted means of predicted adjusted pre-
 200 cipitation accumulations and precipitable water over all ensemble members and all forecast grid
 201 points in $N(s)$. The third statistic measures the dispersion of the predicted precipitation accumula-
 202 tions both between ensemble members and between grid points in $N(s)$. Unlike Scheuerer (2014),
 203 we do not use separate measures of dispersion for those two sources of variability in order to keep
 204 the number of parameters in our heteroscedastic regression model (defined below) as few as possi-

ble. We finally note that the adjustment of forecasts in $N(s)$ to the observation climatology at s via quantile mapping achieves two goals: first, it produces an implicit downscaling to the precipitation at elements on a finer grid; second, it results in a homogenization of the forecasts within $N(s)$, so that the aggregation of forecasts within a large neighborhood to ensemble statistics is reasonable also in, e.g. mountainous regions with substantially varying climatologies.

b. Unconditional precipitation accumulations

Although our main interest is in modeling the conditional distribution of observed precipitation accumulations given the ensemble forecasts, we first consider their unconditional (i.e. climatological) distributions. Studying those is much easier and yet quite instructive, as the conditional distributions should converge towards the unconditional distribution as forecast skill decreases. Moreover, they will allow us to parameterize the conditional distributions such as to make them more comparable across grid points with different climatologies. To fit the parametric CDF $\tilde{F}_{\mu,\sigma,\delta}$ to the empirical CDF \hat{F}_n of the observations y_1, \dots, y_n at this grid point, we minimize the integrated quadratic distance

$$d_{IQ}(\tilde{F}_{\mu,\sigma,\delta}, \hat{F}) = \int_0^\infty (\tilde{F}_{\mu,\sigma,\delta}(t) - \hat{F}_n(t))^2 dt \quad (7)$$

in μ , σ , and δ . According to Thorarinsdottir et al. (2013), this is equivalent to minimizing the mean continuous ranked probability score (CRPS)

$$\frac{1}{n} \sum_{i=1}^n \text{crps}(\tilde{F}_{\mu,\sigma,\delta}, y_i) \quad (8)$$

where

$$\text{crps}(F, y) = \int_{-\infty}^{\infty} (F(t) - H(t - y))^2 dt, \quad (9)$$

222 and $H(\cdot)$ is the Heaviside step function, i.e. it is equal to 1 if $t \geq 0$ and zero otherwise. After
 223 re-parameterizing, the integral on the right hand side can be expressed in closed form as

$$\begin{aligned} \text{crps}(\tilde{F}_{k,\theta,\delta}, y) &= \theta \tilde{y} (2F_k(\tilde{y}) - 1) - \theta \tilde{c} F_k(\tilde{c})^2 \\ &\quad + \theta k (1 + 2F_k(\tilde{c})F_{k+1}(\tilde{c}) - F_k(\tilde{c})^2 - 2F_{k+1}(\tilde{y})) \\ &\quad - \frac{\theta k}{\pi} B\left(\frac{1}{2}, k + \frac{1}{2}\right) (1 - F_{2k}(2\tilde{c})) \end{aligned}$$

224 where $\tilde{c} := -\frac{\delta}{\theta}$, $\tilde{y} := \frac{y-\delta}{\theta}$ and $B(\cdot, \cdot)$ is the beta function (a derivation of this formula is given
 225 in online appendix B). The availability of a closed form expression makes model fitting through
 226 numerical CRPS minimization computationally efficient. When performing this minimization, the
 227 constraint $\delta \geq -\mu$ is imposed in addition to the constraints $\mu, \sigma > 0$ and $\delta \leq 0$ that are required for
 228 the distribution model to be well-defined. The reason for this will become more clear later in this
 229 section, when we set up the regression model for the conditional distribution of the observation
 230 given the forecasts.

231 For solving the constrained optimization problems numerically, we use the Fortran 77 imple-
 232 mentation of the Linearly Constrained Optimization Algorithm (LINCOA) by Michael J. D. Pow-
 233 ell (details of this algorithm have not been published yet, but the usual way of choosing a new
 234 vector of variables is described in Powell 2014). A starting value for the optimization is obtained
 235 through the following rationale: if we had $\sigma = \mu$, the underlying gamma distribution would have
 236 a shape parameter $k = 1$, which corresponds to the special case of an exponential distribution. For
 237 this distribution, the mean over all non-zero precipitation amounts is an estimate of μ (and σ),
 238 for any probability of precipitation π_{pop} , and δ can subsequently be estimated as $\delta = \mu \log(\pi_{pop})$.
 239 For the 12h-accumulations considered here, the best-fitting k is typically smaller than 1, with μ
 240 being overestimated by the assumption of an exponential distribution. Moreover, the first guess
 241 estimates proposed above might violate the constraint $\delta \geq -\mu$. We therefore improve our first

242 guess by fixing σ , gradually decreasing μ , and recalculating $\delta = -(\mu/k) \cdot F_k^{-1}(1 - \pi_{pop})$ until
 243 $\delta > -\mu/2$. The resulting values of μ , σ , and δ are then used as starting values for the numerical
 244 minimization of (8). If $\pi_{pop} < 0.02$, we expect the number of days with non-zero precipitation
 245 to be too small to warrant stable estimates, and we therefore take the starting values as the final
 246 estimates. For extremely dry grid points with $\pi_{pop} < 0.005$, even the simple preliminary esti-
 247 mates might be unreliable, and we use ad-hoc values $\mu = 0.0005$, $\sigma = 0.0182$, $\delta = -0.00049$ to
 248 set up a parametric distribution model for the analyzed climatology. This choice complies with
 249 the constraint $\delta \geq -\mu$ and corresponds to a CSGD distribution with a probability of precipitation
 250 of slightly less than 0.005.

251 Figs. 3 and 4 show examples of fitted CSGDs at a very wet location (West Palm Beach, FL) and
 252 a very dry location (Phoenix, AZ), respectively. The empirical and the fitted, parametric CDFs
 253 are virtually indistinguishable. The approximate character of the parametric distribution becomes
 254 more obvious when we plot its quantiles against the sorted observations. In those Q-Q plots we
 255 observe quite strong departures from the diagonal, especially in the upper tail. However, this is
 256 also where we expect significant sampling variability. In order to understand to what extent the de-
 257 partures might just be random, we add pointwise 95% Monte Carlo intervals by simulating 10000
 258 samples of the same size as the original observations according to the fitted distribution model,
 259 sorting them, and reporting the 2.5% and 97.5% quantile of the first elements, second elements,
 260 and so forth. The black dots in the Q-Q plots in Fig. 3 and 4 (and in all other examples that we
 261 studied) are mostly inside the 95% Monte Carlo intervals, suggesting that the distribution family
 262 proposed here is adequate for modeling unconditional distributions of precipitation accumulations.

c. Regression equations

The final step is now to set up and fit a regression model for the conditional distribution of observed precipitation accumulations given the forecasts. To this end, the ensemble statistics for location s defined above must be linked to the parameters μ_s , σ_s , and δ_s of our CSGD model in eqs. (1) and (2). Denote by $\mu_{cl,s}$, $\sigma_{cl,s}$ and $\delta_{cl,s}$ the parameters of the climatological CSGD at s , and by $\bar{f}_{cl,s}$ and $\bar{\chi}_{cl,s}$ the climatological means of \bar{f}_s and $\bar{\chi}_s$, respectively, calculated as averages of these quantities over the current training sample. We fix $\delta_s = \delta_{cl,s}$ and model the conditional CSGDs as deviations from the climatological CSGD via the following equations

$$\mu_s = \frac{\mu_{cl,s}}{\alpha_{1,s}} \cdot \log 1p \left\{ \exp m1(\alpha_{1,s}) \left[\alpha_{2,s} + \alpha_{3,s} \text{POP}_{f,s} + \alpha_{4,s} \frac{\bar{f}_s}{\bar{f}_{cl,s}} + \alpha_{5,s} \frac{\bar{\chi}_s}{\bar{\chi}_{cl,s}} \right] \right\} \quad (10)$$

$$\sigma_s = \alpha_{6,s} \cdot \sigma_{cl,s} \cdot \left(\frac{\mu_s}{\mu_{cl,s}} \right)^{\alpha_{7,s}} + \alpha_{8,s} \cdot \text{MD}_{f,s} \quad (11)$$

where $\log 1p(x) = \log(1+x)$ and $\exp m1(x) = \exp(x) - 1$. The form of these regression equations, which depend on the fitted parameters $\alpha_{1,s}, \dots, \alpha_{8,s}$, requires some explanation. Consider first a situation with very high predictability. In this case, we typically have $0 < \alpha_{1,s} \ll 1$, which implies $\exp(\alpha_{1,s}) \approx 1 + \alpha_{1,s}$ and $\log(1 + \alpha_{1,s}) \approx \alpha_{1,s}$, and thus eq. (10) reduces to a linear regression equation

$$\mu_s = \mu_{cl,s} \left[\alpha_{2,s} + \alpha_{3,s} \text{POP}_{f,s} + \alpha_{4,s} \frac{\bar{f}_s}{\bar{f}_{cl,s}} + \alpha_{5,s} \frac{\bar{\chi}_s}{\bar{\chi}_{cl,s}} \right] \quad (12)$$

Some exploratory analysis (see also Fig. 6), however, suggests that a linear increase of μ_s with \bar{f}_s is not always appropriate. In situations with reduced predictability (e.g., longer lead times, warm season), ensemble forecasts of high precipitation amounts are particularly unreliable and should be decreased proportionately more compared to forecasts of intermediate levels. This is the rationale behind the logarithm in eq. (10). Increasing the parameter $\alpha_{1,s}$ reduces the growth of μ_s with increasing predictors and thus accounts for the phenomenon just described. Eq. (11) accounts for the heteroscedasticity in the uncertainty about precipitation accumulations in two dif-

283 ferent ways. The first term increases σ_s proportionally to a power of μ_s , which accounts for the
 284 fact that forecast uncertainty increases with the magnitude of expected precipitation amounts. The
 285 second term is proportional to $MD_{f,s}$ and thus accounts for flow-dependent uncertainty. Other
 286 parametric post-processing methods for precipitation (e.g. extended logistic regression) deal with
 287 heteroscedasticity by applying power-transformations to both forecasts and observations with the
 288 goal of making their relation more homoscedastic. This might be preferable when only small
 289 training datasets are available because it results in a potentially less complex model for σ_s . It
 290 entails, however, the disadvantage of strongly distorting the scale of these variables. Consider,
 291 for example, the two hypothetical five member ensembles $(0.5, 1, 1.5, 2, 10)$ and $(0, 2.5, 3.5, 4, 5)$
 292 which have the same mean, but mean absolute differences of 4.0 and 2.3, respectively. The higher
 293 dispersion of the first ensemble results from one member predicting a substantially higher precip-
 294 itation amount than the other members which indicates a certain chance for heavier precipitation.
 295 If the mean absolute differences were calculated from cube root transformed forecasts, values of
 296 0.595 and 0.730 would be obtained, suggesting more uncertainty in the second ensemble. This
 297 does not adequately reflect the situation in the original ensembles, and could thus reduce the value
 298 of flow-dependent uncertainty information in the ensemble. Modeling heteroscedasticity explic-
 299 itly as in eq. (11) avoids the need for data transformations but entails a more complex regression
 300 model. Including $\mu_{cl,s}$ and $\sigma_{cl,s}$ in the two regression equations does not change the actual model
 301 but is useful because it normalizes the regression parameters $\alpha_{1,s}, \dots, \alpha_{8,s}$ and makes them more
 302 comparable across grid points with different climatologies.

303 Fig. 5 illustrates the evolution of the predictive CSGD density with increasing mean precipita-
 304 tion \bar{f}_s in a simplified setting where $\alpha_{3,s}, \alpha_{4,s}$ and $\alpha_{8,s}$ have been set to zero. It shows how the
 305 uncertainty increases with increasing \bar{f}_s ; at the same time the skewness of the underlying gamma

306 distribution becomes smaller and smaller. Choosing $\alpha_{1,s} = 0.05$ results in a moderate departure
 307 from a linear relation between \bar{f}_s and μ_s .

308 Is the CSGD adequate for modeling conditional distributions of precipitation accumulations,
 309 and are the above regression equations for its parameters μ and σ adequate for describing the evo-
 310 lution of these parameters with increasing ensemble mean? To answer this we compare quantiles
 311 derived from predictive CSGDs with empirical conditional quantiles obtained without any para-
 312 metric assumption. For this purpose, however, even the 12 years worth of reforecast data are not
 313 enough if only data from a single grid point are considered. We focus on the analysis grid point
 314 corresponding to the city of Atlanta, GA, and we increase the corresponding dataset by selecting
 315 200 additional analysis grid points within a radius of about 700 km around Atlanta that have a sim-
 316 ilar climatology and are at least 40 km apart from each other. For each season, we then have about
 317 $91 \times 12 \times 201$ pairs of observations and quantile adjusted forecasts. We study again the simplified
 318 regression model with $\alpha_{3,s} = \alpha_{4,s} = \alpha_{8,s} = 0$, i.e. with \bar{f}_s as the only predictor. The conditional
 319 quantiles of the observation given $\bar{f}_s = x$ can then be approximated by considering all forecast-
 320 observation pairs for which \bar{f}_s falls within a certain window $(x - \varepsilon, x + \varepsilon)$ around the precipitation
 321 amount x , and computing the quantiles of the corresponding observations. We let ε increase with
 322 x to account for the fact that the number of pairs with $\bar{f}_s \approx x$ decreases rapidly as x increases.
 323 For $x = 5$ mm and $x = 15$ mm our choice of ε is illustrated in Fig. 6. The crosses in each plot
 324 correspond to the empirical, conditional deciles (i.e. quantiles for the probabilities $0.1, \dots, 0.9$)
 325 for each season and forecast lead times $+12$ to $+24$ h and $+108$ to $+120$ h. The solid lines are
 326 the quantiles obtained with our parametric regression model, fitted to the same training data. As
 327 for the unconditional CSGDs, the regression parameters are fitted by CRPS minimization using
 328 the LINCOA algorithm. Clearly, not every model-based quantile approximates the respective em-
 329 pirical quantile perfectly, and very irregular behavior cannot be captured. Yet one can see that the

the non-linear relation between \bar{f}_s and μ_s , which takes different forms depending on season and lead time, is accounted for by the logarithm in eqn. (10), allowing the red median curves to bend downward from a linear curve as \bar{f}_s increases. Moreover, the increase of predictive uncertainty (distances between the blue decile curves) with increasing \bar{f}_s is captured quite well by the model for σ_s given in eqn. (11). It is worth noting that our method for getting empirical estimates of conditional quantiles is quite similar to what is being done by analog approaches. Those techniques are much more flexible and avoid the approximation errors entailed by parametric methods. On the other hand, several of the plots in Fig. 6 also suggest that the empirical quantiles for large values of \bar{f}_s are subject to quite substantial sampling error, even in the situation considered here where we choose the “analog” from a training data set of size $91 \times 12 \times 201$.

Finally, consider how the regression model (10), (11) for the predictive CSGDs approaches the parameters for the climatological CSGD in the limit where the raw ensemble forecasts have no skill. As the lead time increases, one can expect that the four predictors $\text{POP}_{f,s}$, \bar{f}_s , $\bar{\chi}_s$ and $\text{MD}_{f,s}$ become less and less informative about the true weather, and so the corresponding regression parameters $\alpha_{3,s}$, $\alpha_{4,s}$, $\alpha_{5,s}$, and $\alpha_{8,s}$ tend to zero. If at the same time $\alpha_{2,s}$ and $\alpha_{6,s}$ tend to one, then μ_s and σ_s tend to the climatological CSGD parameters $\mu_{cl,s}$ and $\sigma_{cl,s}$, whatever the values of $\alpha_{1,s}$ and $\alpha_{7,s}$, and so the climatological CSGD results as a limiting case. Including $\mu_{cl,s}$ and $\sigma_{cl,s}$ in the regression equations (10), (11) can therefore be seen as a kind of normalization which helps reduce the dependence of the regression parameters $\alpha_{1,s}, \dots, \alpha_{8,s}$ on the climatology at location s and thus renders them more comparable across different gridpoints.

Modeling the conditional distributions as deviations from the climatological distributions requires some constraints of the latter. We found that this deviation concept does not work well at very dry locations if the shift parameter $\delta_{cl,s}$ of the climatological CSGD is large compared to $\mu_{cl,s}$. In this case, positive precipitation accumulations correspond to the tail end of the under-

354 lying gamma distribution, and deforming this distribution into a CSGD with a moderate to high
 355 probability of precipitation is rather unnatural. By introducing the constraint $\delta_{cl,s} \geq -\mu_{cl,s}$ on the
 356 climatology parameters in subsection b), we enforce a very small shape parameter k . The mass
 357 of the underlying gamma distribution is then concentrated near zero, and a very small shift is suf-
 358 ficient to obtain a high probability of values less than zero. Fitting a climatological CDF to the
 359 observation data under this constraint can result in a slightly sub-optimal fit to the empirical, cli-
 360 matological CDF near zero, but this degradation is offset by the fact that the fitted CSGD permits
 361 a natural deformation into the predictive CSGD for any value of the predictors.

362 **5. Validation of the CSGD method**

363 We apply our CSGD regression method to the full data set described in Section 2. Now, every
 364 grid point of the CCPA grid (within the CONUS) is processed separately. Forecasts are cross-
 365 validated; for example, 2002 forecasts are trained using 2003-2013 data. In order to account for
 366 seasonal differences, a separate set of (both climatological and regression) parameters is fitted for
 367 each month; training data is composed of all forecasts and observations from ± 45 days around
 368 the 15th of the month under consideration and all years except the one for which forecasts are
 369 sought. This results in a training sample size of 91×11 at each grid point. Compared to the
 370 amount of training data that is typically used for weather variables like wind speed or temperature,
 371 this training sample size seems fairly large. At very dry locations, however, the majority of both
 372 forecasts and observations are zero, and thus carry only limited information that can be leveraged
 373 for model fitting. For the parameters of the unconditional CSGDs we already described our special
 374 treatment of these dry cases in Section 4b. For the regression parameters, we increase the training
 375 data set of any grid point where the climatological probability of precipitation is less than 0.05 by
 376 considering also the data at adjacent grid points in east-west and north-south direction. For grid

377 points with a climatological probability of precipitation of less than 0.02, we additionally add the
 378 training data from diagonal neighbors. Parameters are estimated via CRPS minimization, subject
 379 to the following bounds:

$$0.001 \leq \alpha_{1,s}, \alpha_{2,s} \leq 1 \quad , \quad 0 \leq \alpha_{3,s}, \alpha_{4,s}, \alpha_{5,s} \leq 1.5,$$

$$0.1 \leq \alpha_{6,s}, \alpha_{7,s} \leq 1 \quad , \quad 0 \leq \alpha_{8,s} \leq 1.5,$$

380 which are partly ad hoc and partly based on the discussion at the end of the previous section.
 381 In our experiments, CRPS minimization gave slightly better results than classical maximum
 382 likelihood estimation, which is nonrobust and tends to favor over-dispersive predictive CSGDs.
 383 The same conclusion was reached by Gneiting et al. (2005) in the context of temperature post-
 384 postprocessing. Initially, we fix the radius of the neighborhood within which forecasts are consid-
 385 ered as predictors (see Section 4a) to $r = 2$ degrees (≈ 200 km).

386 *a. Overall performance and model complexity*

387 First, we take a look at the overall predictive performance of our CSGD method, measured by the
 388 continuous ranked probability skill score (CRPSS), which quantifies the improvement of the CRPS
 389 of the predictive CSGDs over climatological forecasts. We also study in how far the different
 390 nonlinear and heteroscedastic components of our model contribute to this overall performance. To
 391 this end we consider five submodels of the full regression model (10), (11):

- 392 1. basic linear model (12) for μ_s with the ensemble mean as the only predictor (i.e. $\alpha_{1,s} = \alpha_{3,s} =$
 393 $\alpha_{5,s} = 0$); increase of σ_s proportional to $\sqrt{\mu_s}$ and no use of $MD_{f,s}$ (i.e. $\alpha_{7,s} = 0.5, \alpha_{8,s} = 0$).
- 394 2. as the previous model but with the nonlinear model (10) for μ_s instead of (12).
- 395 3. as the previous model but releasing the ad-hoc assumption $\alpha_{7,s} = 0.5$ about the rate of increase
 396 of σ_s with increasing μ_s .

397 4. as the previous model but with added predictor $MD_{f,s}$ for σ_s .

398 5. as the previous model but with added predictor $\bar{\chi}_s$ (precipitable water) for μ_s .

399 The basic model has only three parameters and is comparable (in terms of model complexity) with
400 the (basic) extended logistic regression model (Wilks 2009). Direct comparisons of the parametric
401 post-processing approaches mentioned in Section 1 (ExLR, BMA, EMOS) suggest that their pre-
402 dictive performance is quite similar (Schmeits and Kok 2010; Scheuerer 2014), so how much extra
403 skill can be gained by adding additional predictors or permitting certain forms of nonlinearity?

404 Fig. 7 depicts the overall CRPSS (for the full model) for different lead times and the CRPSS
405 increase that results from adding step-by step the extra components described above. The first
406 thing to note is the pronounced seasonal cycle of the CRPSS. Summertime convection is more
407 difficult to forecast than synoptic-scale winter precipitation, and so forecast skill during the cool
408 season is substantially higher than during the warm season. This pattern is inherited from the raw
409 ensemble predictions, the corresponding results can be found in Hamill et al. (2015). The increase
410 in skill due to the different refinements of the basic model is rather moderate for each individual
411 extension, but sums up to a cumulative increase of about 0.01 to 0.015. The biggest benefit results
412 from allowing a nonlinear increase of μ_s with \bar{f}_s , especially for longer lead times (see right panel
413 of Fig. 7). The predictor $POP_{f,s}$ yields a rather constant improvement in skill over all months of
414 the year, while the predictor $\bar{\chi}_s$ (precipitable water) becomes especially useful in the warm sea-
415 son but adds no information to the ensemble precipitation forecasts during the more predictable
416 cool season. The converse is true for the $MD_{f,s}$ predictor which measures the spread of the fore-
417 casts between different ensemble members and forecast grid points within $N(s)$: it provides useful
418 information about flow-dependent forecast uncertainty during the cool season, but does not im-
419 prove (or even degrades, for longer lead times) probabilistic forecast skill during the warm season.

420 The degradation is presumably a result of overfitting, to which the $MD_{f,s}$ predictor is particularly
421 prone, and which becomes a more serious concern as the signal to noise ratio in the training dataset
422 decreases. Finally, we note that estimating the rate of increase of σ_s with increasing μ_s rather than
423 fixing $\alpha_{7,s} = 0.5$ adds some flexibility, but the resulting benefit on predictive performance is quite
424 marginal.

425 Fig. 8 depicts maps of CRPSS values of the CSGD predictions to provide an impression about
426 regional differences. Forecast skill is largest along the east and especially the west coast of the
427 CONUS, which we believe is due to the relatively high predictability of orographically induced,
428 synoptically forced precipitation. The general spatial pattern of forecast skill resembles that of the
429 raw ensemble (see Hamill et al. 2015) while the skill is significantly better.

430 *b. CSGD vs. analog method: Brier skill scores and reliability*

431 The CRPS studied so far is a useful and common measure of the overall skill, but it does not
432 allow any conclusion about how skillful the CSGD forecasts are for predicting light, intermediate,
433 and heavy precipitation events. To answer this question, we study Brier skill scores (Wilks 2011,
434 eqs. 7.34 and 7.35) for the three thresholds of 1, 10, and 25 mm 12 h⁻¹. We further compare the
435 predictive performance of the CSGD approach to a recently proposed variant of the rank analog
436 approach by Hamill and Whitaker (2006), where supplemental location are used to augment the
437 training dataset at each analysis grid point. This adjustment to the rank analog procedure was
438 shown to substantially improve probabilistic forecasts for heavy precipitation events (Hamill et al.
439 2015). Can the same or even more improvement be achieved by a parametric post-processing
440 scheme? Fig. 9 depicts the monthly Brier skill scores (BSSs) for both methods, the three different
441 thresholds, and forecast lead times up to +6 days. Even for the >1 mm 12 h⁻¹ event, the CSGD
442 method can still improve upon the analog method, despite the fact that this is a rather common

event at most grid points, and we should expect that sufficiently close analogs can usually be found. The fact that the CSGD method can compete with the analog approach in this situation suggests that our parametric approximation does not degrade predictive performance even when analog methods can be expected to perform very well. Comparing results for higher thresholds, we find that the probabilistic CSGD forecasts are again able to improve upon the forecasts by the analog method. The event $>25 \text{ mm } 12 \text{ h}^{-1}$ is relatively rare, making it difficult to find a sufficient number of suitable analogs, even if supplemental locations are added to increase the training data sets. Our parametric method, on the contrary, can extrapolate relations found for more common situations and thus yield superior predictions of rare events.

To provide some understanding about the causes of the better performance of our parametric method compared to the non-parametric analog approach, we consider reliability diagrams for the same events as above (thresholds 1, 10, and $25 \text{ mm } 12 \text{ h}^{-1}$) and lead times +12 to +24 h and +108 to +120 h. The plots in Figs. 10 and 11 suggest that both methods yield reliable probabilistic forecasts at short lead times. At longer lead times, they are still sufficiently accurate, though somewhat less reliable. By comparing the inset frequency histograms, one can see that the performance gain of our CSGD method is mainly due to increased resolution; it issues high probabilities for observing heavy precipitation more frequently without degrading the reliability compared to the analog approach, which does not rely on parametric assumptions.

We illustrate the last point by considering a heavy precipitation event that took place over Washington state between 1200 UTC on November 6 and 0000 UTC on November 7 in 2006. Fig. 12 shows the analyzed precipitation accumulations for that period, as well as +12 to +24 h lead predicted probabilities for exceeding $25 \text{ mm } 12 \text{ h}^{-1}$ of precipitation by the raw ensemble, the analog approach and the CSGD regression method. The raw ensemble forecasts for that day were quite accurate, but since this is not always the case, one can expect that calibrated probabilistic

forecasts modulate the high forecast probabilities. The analog approach modulates them more strongly, issuing rather moderate probabilities. On the other hand, the CSGD method largely retains the strong signal from the raw ensemble, and hence provides decision makers with a more unequivocal expectation of heavy precipitation.

c. Performance with a greatly reduced training data set

The results by Hamill et al. (2015) underscore the importance of a sufficiently large training data set for statistical post-processing, especially when the interest is in heavy precipitation events. What if a large reforecast dataset is not available? Can the CSGD approach retain its strong performance, or will it lose a large amount of skill as a result of over-fitted regression parameters? Can possible over-fitting be avoided by supplementing the training dataset at each grid point with training data from other grid points? To answer these questions we repeat the entire procedure (quantile mapping, calculating ensemble statistics, fitting the regression model) described above, this time using, for each of the 12 verification years, only forecast data from the preceding year or from the preceding three years (defining 2013 to be the year that precedes 2002) for training. We are thus left with only 91 and 273 training days, respectively, for quantile mapping and model fitting. Using just a single year of training data mimics the situation where no reforecasts are produced, but one year of training data is available from a pre-operational test phase after a major update of the NWP system. Such update would only have a limited or no effect on the verification/calibration data, and we therefore use the same CCPA data sets as before (11 training years for each verification year) for calculating the CCPA quantiles (Section 4a) and fitting the unconditional CSGD model (Section 4b).

Since fixing $\alpha_{7,s} = 0.5$ hardly affected the predictive performance with the large training dataset used above, and since the uncertainty parameter $\alpha_{8,s}$ is particularly prone to overfitting, we fit

reduced CSGD regression models with $\alpha_{7,s} = 0.5$ and $\alpha_{8,s} = 0$ in the present setup. Even the estimation of the remaining six parameters might be difficult with only 91 or 273 training days (the majority of which are typically dry days). We therefore consider a further setup where we use again just one year / three years of training data (forecasts) but supplement the dataset at each analysis grid point with data from 19 other analysis gridpoints with similar climatologies and terrain characteristics, and a certain minimal distance to each other to make sure that their forecast error characteristic are largely independent. For a detailed description of the algorithm for selecting the supplemental locations see Hamill et al. (2015, online appendix A), where supplemental locations are used quite successfully to improve the predictive performance of the analog method for higher precipitation events. In their setup, the supplemental data *complement* the reforecast data; here, we study in how far data from other locations can *substitute* reforecast data.

For the calculation of the forecast quantiles as required for the quantile mapping step, there is no straightforward way to pool data across different grid locations. In this context we must hope that there is sufficient independent information in the ensemble (recall that all ensemble member forecasts are pooled for the purpose of calculating the forecast quantiles) to warrant an adequate estimation of the forecast climatology.

Fig. 13 depicts the decrease of the Brier skill scores obtained with the setups described above compared to the full model fitted with 11 years of training data. The effects of reducing the training sample size are dramatic, especially for the prediction of the 25 mm 12h⁻¹ event. Brier skill scores for +12 to +24 h lead time go down by up to 0.1 when only one year of training data is used. Inspection of the corresponding reliability diagrams (see online appendix C) reveals that the reliability of CSGD forecasts suffers substantially. As a result of overfitting, the predictive CSGDs become overconfident (underdispersive), and this overconfidence particularly affects the higher thresholds. The use of supplemental locations can mitigate but not entirely compensate

the lack of reforecast data; although the training dataset corresponding to the 'one year reforecast plus supplemental locations' setup is almost twice as large as the training dataset with 11 years of reforecasts (but no supplemental data), the resulting CSGD predictions are still inferior. However, they nearly match at least the performance of the CSGD model fitted to a training dataset consisting of three years of reforecasts but no supplemental data. On the one hand this highlights the benefits that a lengthy reforecast can provide, with its greater variety of weather events covered. On the other hand it shows that the strategy of increasing the training dataset by considering supplemental locations can substantially reduce the performance loss due overfitting.

d. Role of the neighborhood size considered for the ensemble statistics

So far, all results for the CSGD method were obtained with a radius $r = 2$ degrees around each analysis grid point, within which forecasts were used as predictors. This is an ad-hoc choice, and the question suggests itself as to how much of an impact the choice of the neighborhood size has on the predictive performance, and what the optimal radius would be for each lead time. To study this, we use again the maximal training dataset (11 years of forecasts and analyses), the full regression model (10), (11), and calculate the CRPSS of the predictive CSGDs for different choices of r . The smallest possible radius $r = 0.5$ degrees (the resolution of the forecast grid) serves as a benchmark and corresponds to neighborhoods that only contain the closest forecast grid points. Extremely large neighborhoods were not tested due to the increased computational expense. In Fig. 14 we depict the change in CRPSS relative to this benchmark value for larger neighborhood sizes. As might be expected, the optimal radius changes with lead time: for the longest (day 4.5 to 5) lead time considered here, the largest radius $r = 3$ yields the best results, while for the shortest (day 0.5 to 1) lead time an initial increase in predictive performance is eventually reversed when r is increased beyond 2 degrees. This case further shows that it is not just lead

time, but more generally predictability that determines the optimal radius: the more predictable precipitation generating processes during the cool season favor smaller neighborhood sizes than the less predictable processes during the warm season. The overall increase of skill resulting from an adequate choice of r (larger than the minimal choice of $r = 0.5$ degrees) is similar or even larger in magnitude than the increase resulting from more sophisticated regression equations as studied in Section 5a.

6. Discussion

We have discussed a parametric post-processing approach that uses statistics of the raw ensemble forecasts as predictors for the parameters of a censored, shifted Gamma distribution (CSGD). Exploratory analysis (see Fig. 3, 4, and 6) showed that CSGDs can approximate both climatological distributions of observed precipitation, and distributions conditional on the ensemble forecasts reasonably well. Ensemble mean and dispersion predictors were estimated from the ensemble at the observation location and in a surrounding area. Ensemble mean precipitable water was used as a further predictor. These statistics were used to drive a heteroscedastic regression model, which was demonstrated to be capable of modeling the relation between ensemble forecasts and parameters of the predictive CSGDs. Verification results showed that the CSGD regression approach yields probabilistic forecast that were sufficiently reliable at all lead times and had better resolution than the forecasts obtained by a state-of-the-art analog approach. This was especially true for forecasts of extreme events, which are of particular interest due to their socio-economic impact.

The CSGD approach presented here adopted the Scheuerer (2014) procedure of utilizing forecasts within a larger neighborhood of the location of interest as predictors. Accordingly, we also studied the connection between the optimal neighborhood size and predictive skill, finding that very large neighborhoods with a radius of over 300 km performed best with longer lead times and

for predictions during the warm season. For short lead times and synoptic-scale winter precipitation, a smaller radius was more appropriate. The improvement in skill with larger neighborhoods compared to a model that only used forecasts at the nearest forecast grid points was similar in magnitude as the improvement due to more complex and flexible regression equations that permitted a nonlinear relation between the predictors and the predictive mean.

Finally, we studied the effect of training sample size on the predictive performance of the fitted CSGD model. For the analog method, a large training data set is preferred because only then can good analogs always be found over all cases. The results presented here suggest that the predictive performance of a parametric approach also suffers substantially if the model is fitted with an insufficiently large data set. Supplemental data from close-by grid points can partly compensate for a lack of reforecasts, but more efficient ways to share information between different locations need to be found to ensure good predictive skill of forecast of more extreme events even with a limited amount of reforecasts. These results affirm the positive value of lengthy training data sets that reforecasts can provide.

References

- Ben Bouallègue, Z., 2013: Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting*, **28**, 515–524.
- Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901.
- Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*,

582 **133**, 1098–1118.

583 Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. G. Jr., Y. Zhu, and
584 W. Lapenta, 2013: NOAA’s second-generation global medium-range ensemble reforecast data
585 set. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565.

586 Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using
587 ECMWF and GFS ensemble reforecasts. part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632.

588 Hamill, T. M., M. Scheuerer, and G. Bates, 2015: Analog probabilistic precipitation forecasts
589 using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, in
590 press.

591 Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on
592 reforecast analogs: theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229.

593 Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: statis-
594 tical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557.

595 Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123**,
596 2181–2196.

597 Messner, J. W., and G. J. Mayr, 2014: Heteroscedastic extended logistic regression for postpro-
598 cessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456.

599 Molteni, F., R. Buizza, T. Palmer, and T. Petroliaigis, 1996: The ECMWF ensemble prediction
600 system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **125**, 73–119.

601 Powell, M. J. D., 2014: On fast trust region methods for quadratic models with linear constraints.
602 Tech. Rep. DAMTP 2014/NA02, Department of Applied Mathematics and Theoretical Physics,
603 Cambridge University.

604 Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model
605 output statistics. *Quart. J. Roy. Meteor. Soc.*, **140** (680), 1086–1096.

606 Schmeits, M. J., and K. J. Kok, 2010: A comparison between raw ensemble output, (modified)
607 Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipi-
608 tation reforecasts. *Mon. Wea. Rev.*, **142**, 4199–4211.

609 Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precip-
610 itation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220.

611 Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl, 2013: Using proper divergence functions to
612 evaluate climate models. *SIAM/ASA J. Uncert. Quant.*, **1** (1), 522–534.

613 Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations.
614 *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

615 Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon.*
616 *Wea. Rev.*, **125**, 3297–3319.

617 Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS
618 forecasts. *Meteor. Appl.*, **16**, 361–368.

619 Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed., Elsevier Academic
620 Press.

621 Wilks, D. S., and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS refore-
622 casts. *Mon. Wea. Rev.*, **135**, 2379–2390.

623	LIST OF FIGURES	
624	Fig. 1.	Examples of censored, shifted gamma distributions. The fractions of the probability density function that fall below zero (shown in the grey shading) translate into a positive probability of being exactly zero. 32
625		
626		
627	Fig. 2.	Illustration of the neighborhood weighting scheme and the climatology adjustment for an analysis grid point ('+') near Sacramento, CA, and $r = 2$ deg. Forecast grid points are denoted by '•', their area is proportional to the weight w_{sx} . The middle and right-hand panel illustrate, for two of these forecast grid points, how the corresponding forecasts are adjusted by quantile mapping. 33
628		
629		
630		
631		
632	Fig. 3.	Empirical and fitted CDFs (top) and Q-Q plots (bottom) of +12 to +24 h analyzed precipitation accumulations in West Palm Beach, FL. The black dots in the lower panels are the sorted observations, plotted against the corresponding theoretical quantiles from the fitted CSGD model. Ideally, they would lie on the diagonal (solid red line); due to sampling variability, however, any black dot lying within the pointwise 95% Monte Carlo intervals (solid blue lines) can still be considered consistent with the fitted model. 34
633		
634		
635		
636		
637		
638	Fig. 4.	Same as Fig. 3, but for Phoenix, AZ. 35
639	Fig. 5.	Example of predictive CSGD densities, showing the evolution of the CSGD parameters μ and σ from eqs. (1) and (2) as a function of the ensemble-mean statistic \bar{f}_s 36
640		
641	Fig. 6.	Conditional deciles (median is highlighted in red) obtained with the augmented Atlanta data set. Empirical deciles are depicted as crosses. For each conditioning value 1 mm, 2 mm, ..., 25 mm they are obtained as empirical deciles of the observations corresponding to ensemble-mean statistics within a certain bin (for 5 mm and 15 mm depicted as vertical dashed lines) around this value. Deciles derived from the CSGD regression model are depicted as solid lines. 37
642		
643		
644		
645		
646		
647	Fig. 7.	CRP skill scores (left panel) for different lead times, separately for each month, but aggregated over all analysis grid points within the CONUS. Increases of CRPSSs due to increased model complexity are shown for for +12 to +24 h lead time (middle panel) and +108 to +120 h lead time (right panel). 38
648		
649		
650		
651	Fig. 8.	Map of CRPSS values, aggregated over all months and all cross-validated years, for +12 to +24 h lead time (left panel) and +108 to +120 h lead time (left panel). 39
652		
653	Fig. 9.	Brier skill scores for different lead times and different event thresholds, separately for each month, but aggregated over all analysis grid points within the CONUS. Results for the rank-analog method are shown in the top row, those for the CSGD regression approach are shown in the bottom row. 40
654		
655		
656		
657	Fig. 10.	Reliability diagrams for +12 to +24 h lead time and different event thresholds, calculated with forecast-observation pairs of all months, all cross-validated years, and all analysis grid points within the CONUS. The top row shows results for the rank-analog method, the bottom row shows results for the CSGD regression approach. The inset histograms depict the frequency with which each category was predicted. 41
658		
659		
660		
661		
662	Fig. 11.	As Fig. 10 but for +108 to +120 h lead time. 42

663	Fig. 12.	Analyzed precipitation between 1200 UTC, Nov 6, 2006 and 0000 UTC Nov 7, 2006 (a)	
664		and corresponding +12 to +24 h lead probability forecasts for exceeding $25 \text{ mm } 12 \text{ h}^{-1}$	
665		of precipitation by the raw ensemble (b), the analog method (c) and the CSGD regression	
666		approach (d).	43
667	Fig. 13.	Decrease of the Brier skill scores (aggregated over all analysis grid points within the	
668		CONUS) due to a reduction of the training sample to one or three years of training data.	
669		In both cases we also give results for CSGD distributions fitted with additional training data	
670		from 19 supplemental locations.	44
671	Fig. 14.	Increase of CRP skill scores for different neighborhood sizes, relative to the smallest possi-	
672		ble neighborhood with $r = 0.5$ degrees.	45

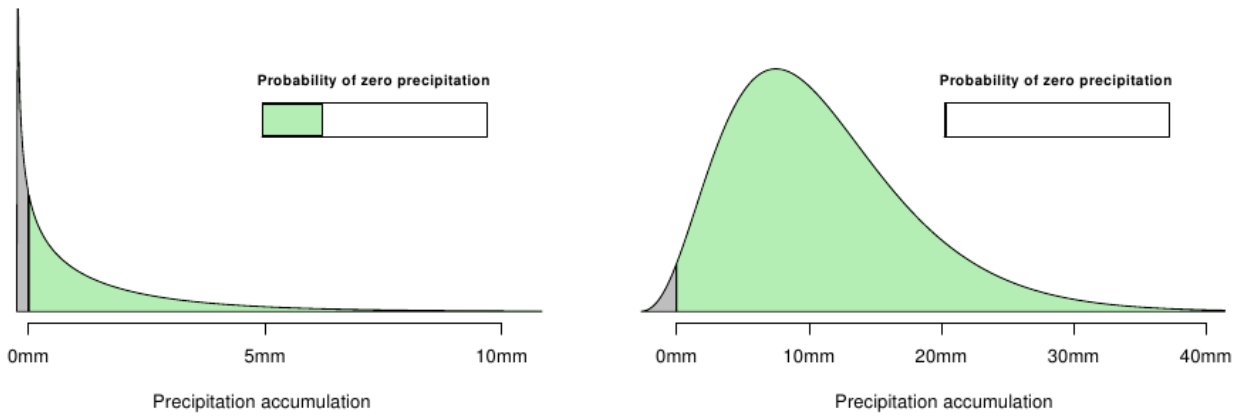


FIG. 1. Examples of censored, shifted gamma distributions. The fractions of the probability density function that fall below zero (shown in the grey shading) translate into a positive probability of being exactly zero.

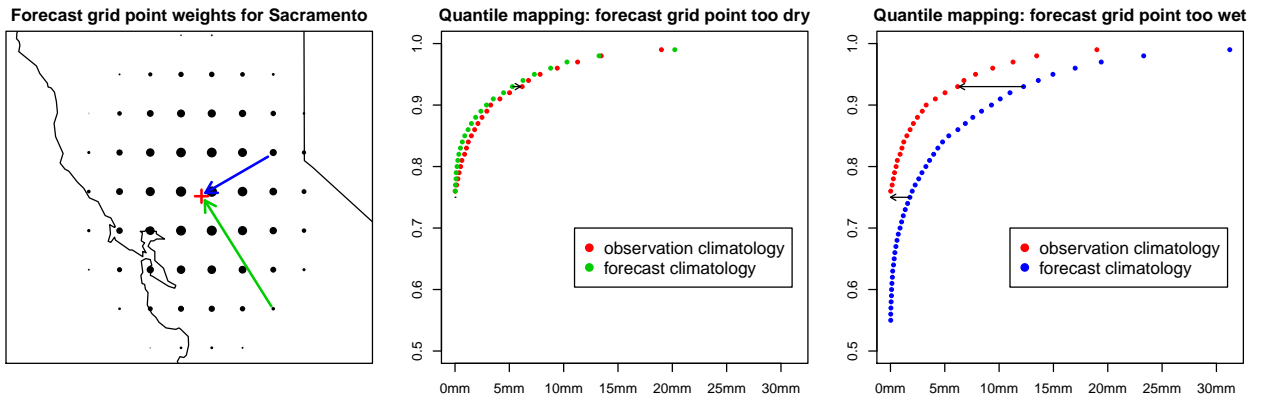


FIG. 2. Illustration of the neighborhood weighting scheme and the climatology adjustment for an analysis grid point ('+') near Sacramento, CA, and $r = 2$ deg. Forecast grid points are denoted by '•', their area is proportional to the weight w_{sx} . The middle and right-hand panel illustrate, for two of these forecast grid points, how the corresponding forecasts are adjusted by quantile mapping.

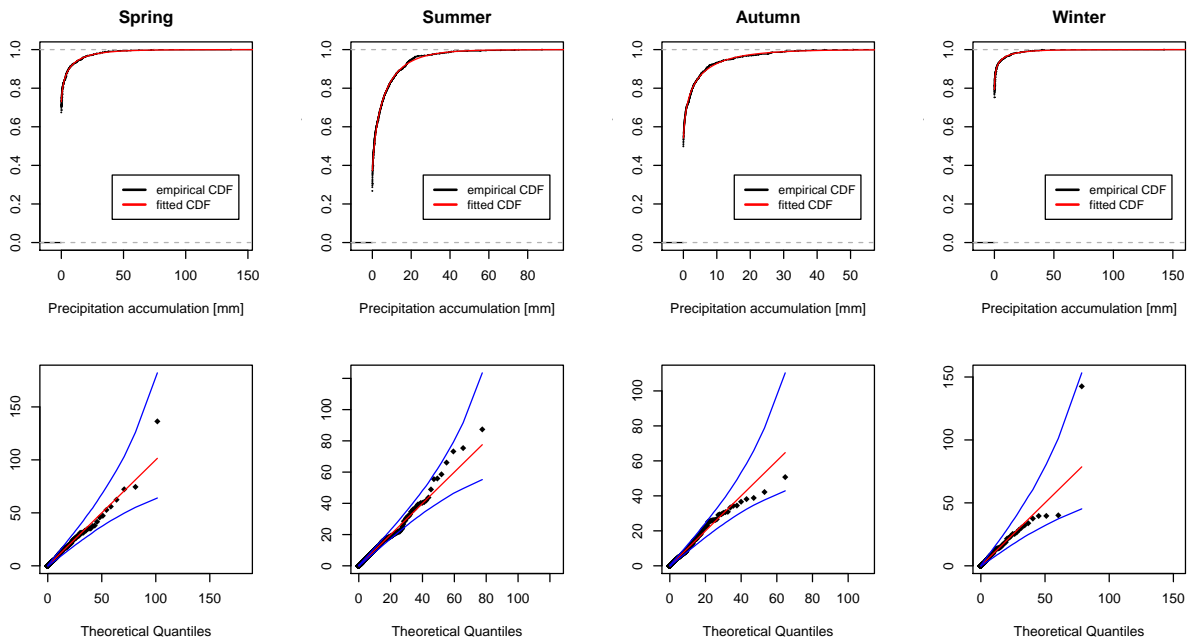


FIG. 3. Empirical and fitted CDFs (top) and Q-Q plots (bottom) of +12 to +24 h analyzed precipitation accumulations in West Palm Beach, FL. The black dots in the lower panels are the sorted observations, plotted against the corresponding theoretical quantiles from the fitted CSGD model. Ideally, they would lie on the diagonal (solid red line); due to sampling variability, however, any black dot lying within the pointwise 95% Monte Carlo intervals (solid blue lines) can still be considered consistent with the fitted model.

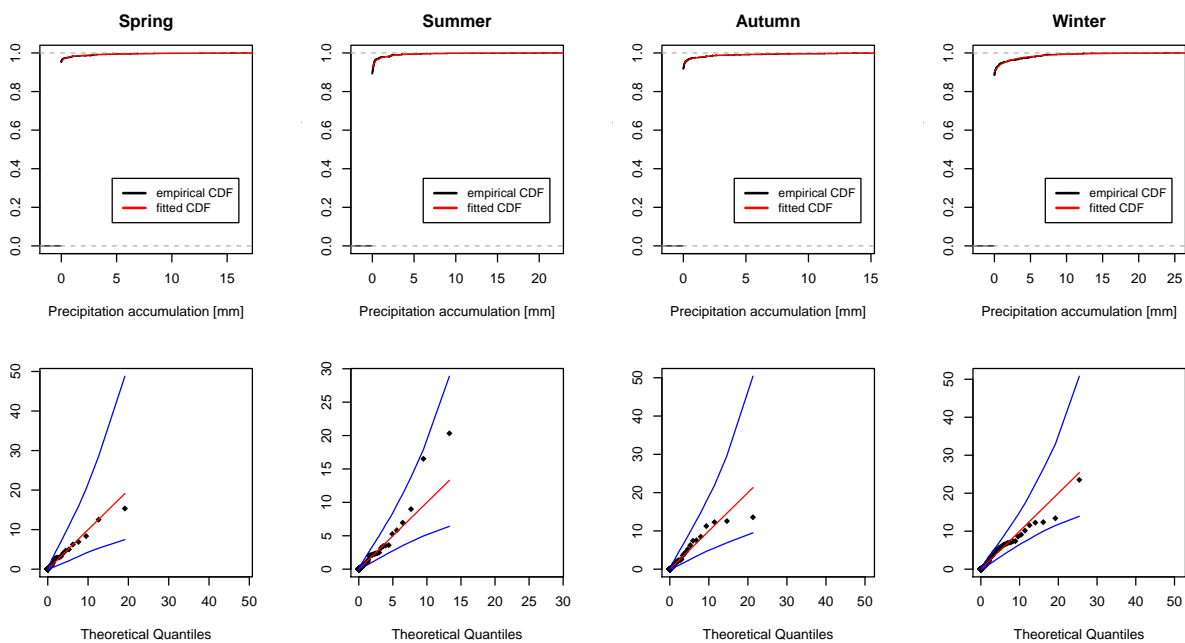


FIG. 4. Same as Fig. 3, but for Phoenix, AZ.

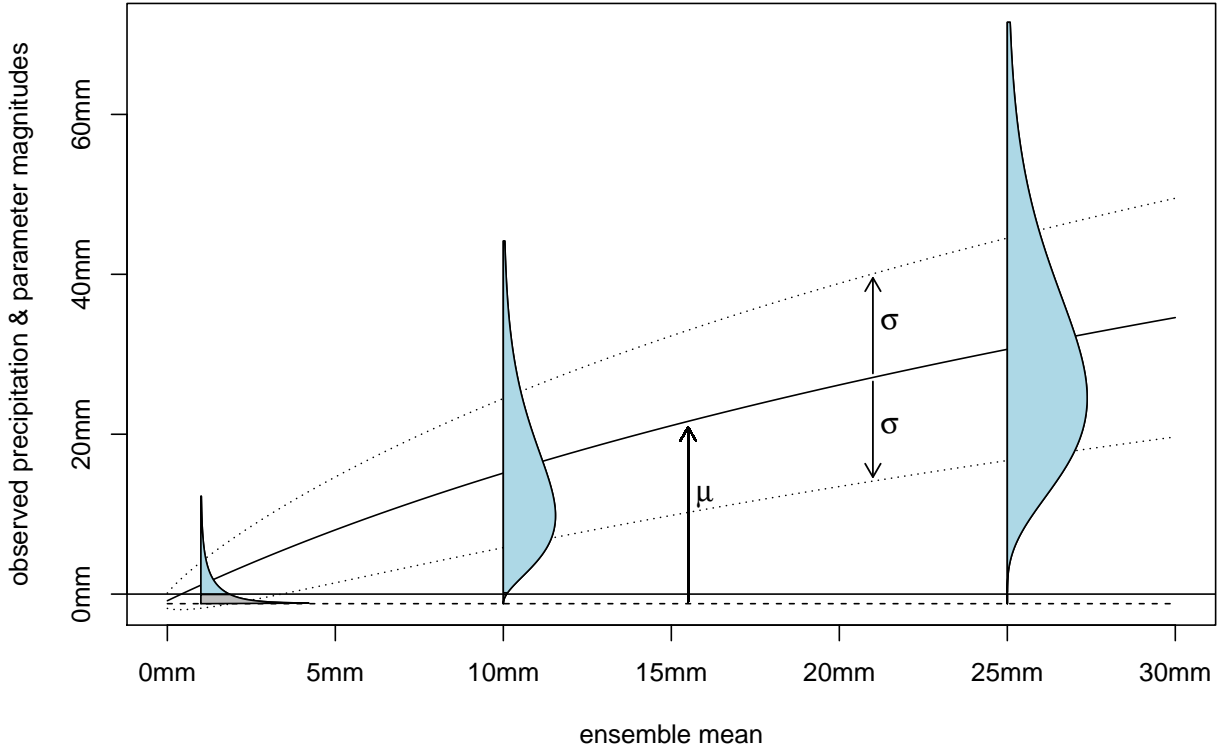


FIG. 5. Example of predictive CSGD densities, showing the evolution of the CSGD parameters μ and σ from eqs. (1) and (2) as a function of the ensemble-mean statistic \bar{f}_s .

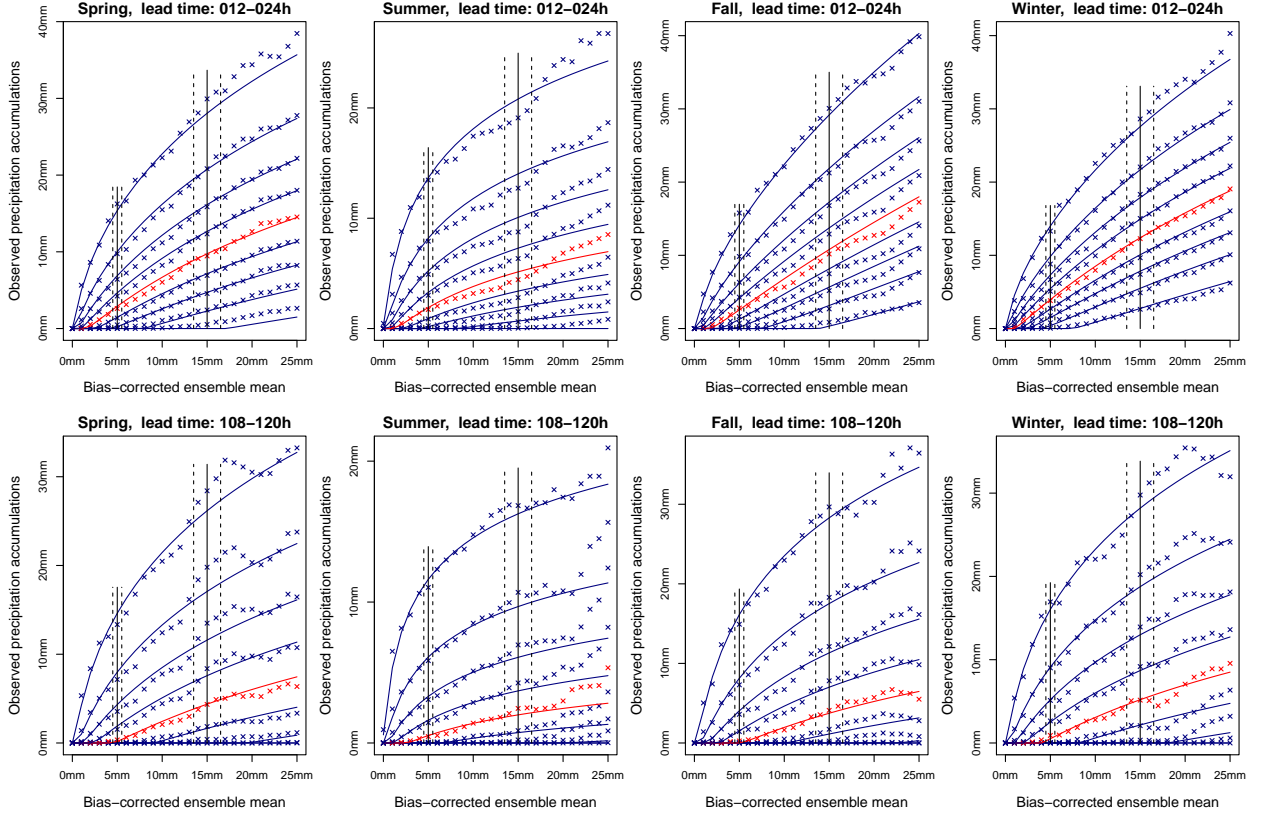


FIG. 6. Conditional deciles (median is highlighted in red) obtained with the augmented Atlanta data set. Empirical deciles are depicted as crosses. For each conditioning value 1 mm, 2 mm, ..., 25 mm they are obtained as empirical deciles of the observations corresponding to ensemble-mean statistics within a certain bin (for 5 mm and 15 mm depicted as vertical dashed lines) around this value. Deciles derived from the CSGD regression model are depicted as solid lines.

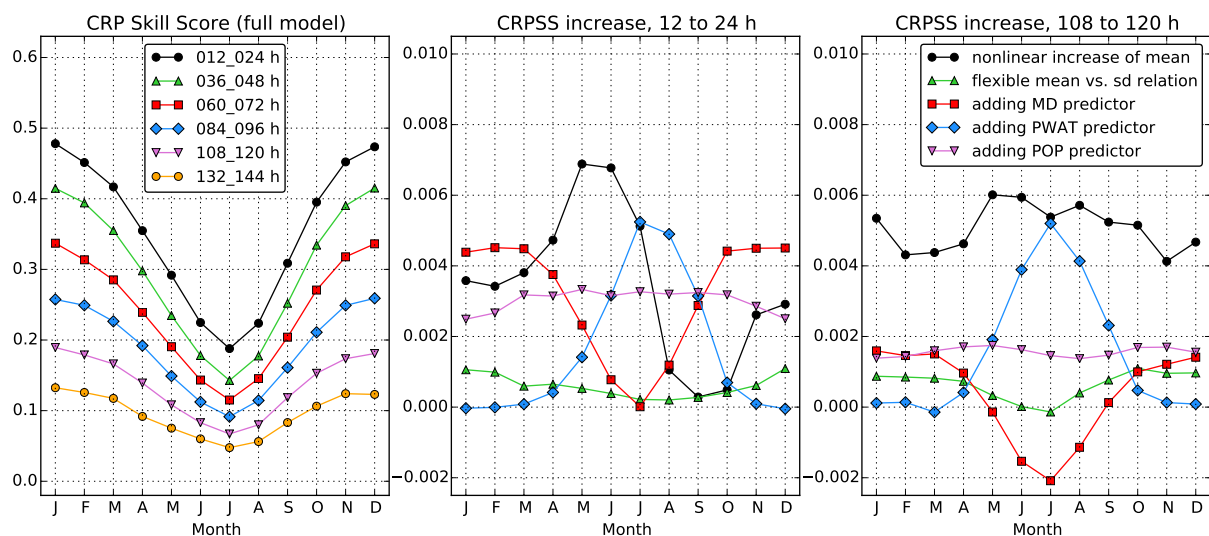


FIG. 7. CRP skill scores (left panel) for different lead times, separately for each month, but aggregated over all analysis grid points within the CONUS. Increases of CRPSSs due to increased model complexity are shown for for +12 to +24 h lead time (middle panel) and +108 to +120 h lead time (right panel).

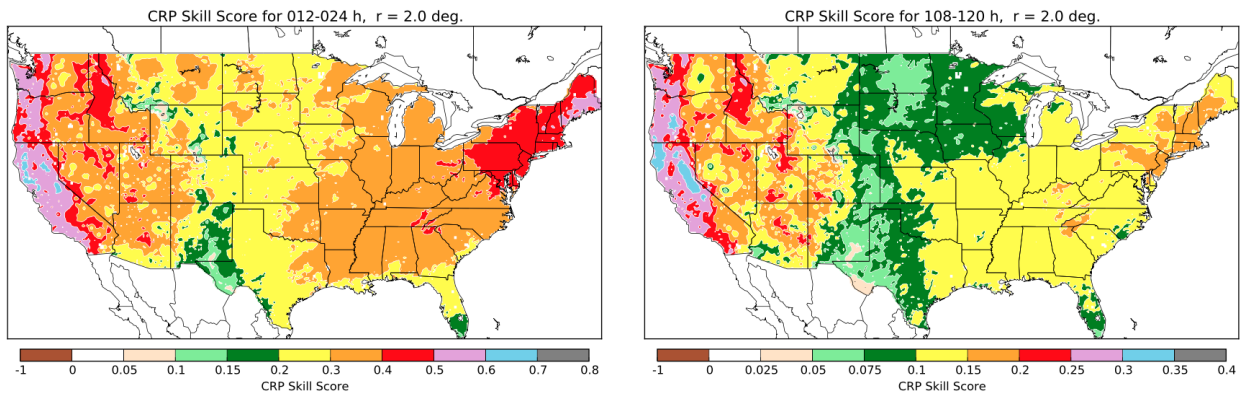


FIG. 8. Map of CRPSS values, aggregated over all months and all cross-validated years, for +12 to +24 h lead time (left panel) and +108 to +120 h lead time (left panel).

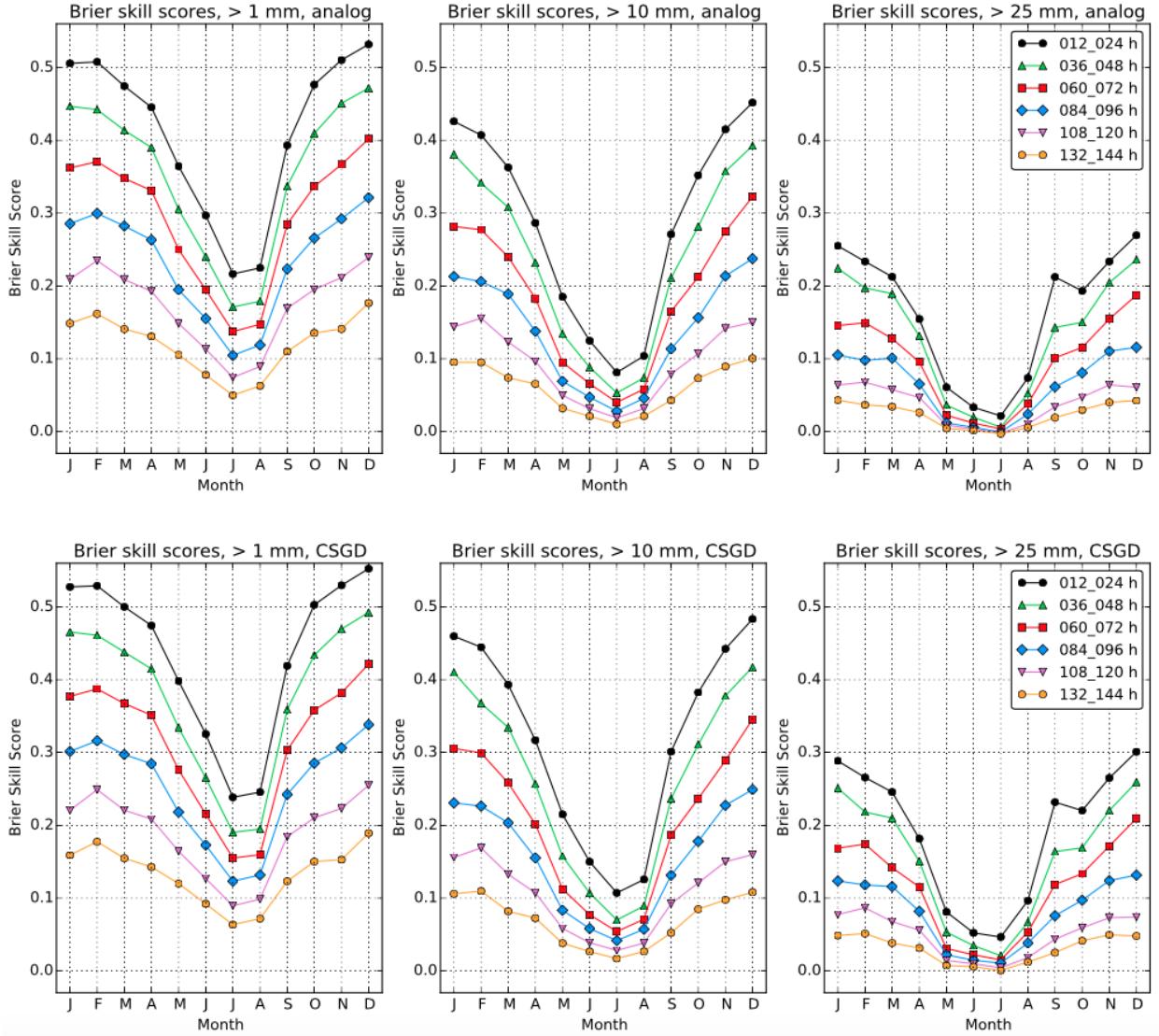


FIG. 9. Brier skill scores for different lead times and different event thresholds, separately for each month, but aggregated over all analysis grid points within the CONUS. Results for the rank-analog method are shown in the top row, those for the CSGD regression approach are shown in the bottom row.

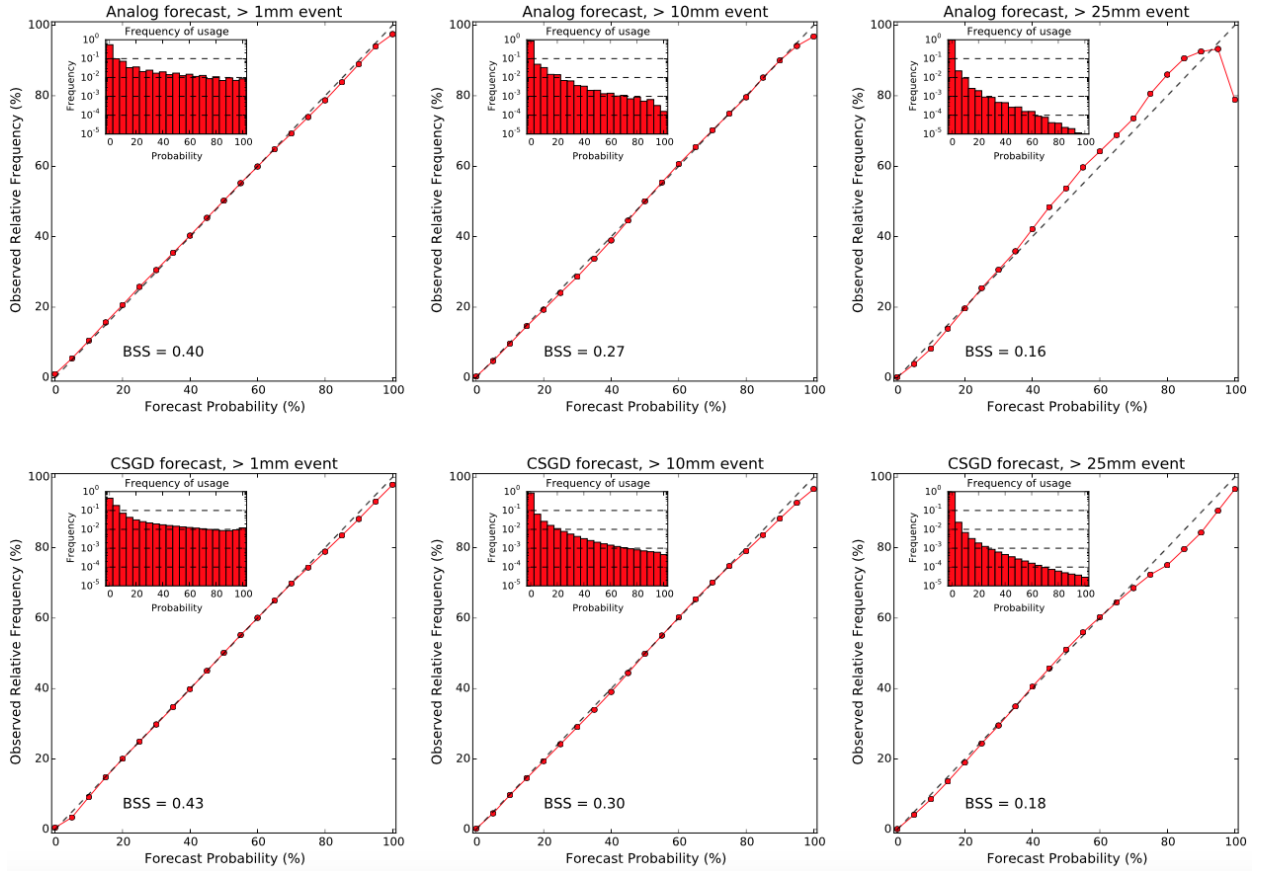


FIG. 10. Reliability diagrams for +12 to +24 h lead time and different event thresholds, calculated with forecast-observation pairs of all months, all cross-validated years, and all analysis grid points within the CONUS. The top row shows results for the rank-analog method, the bottom row shows results for the CSGD regression approach. The inset histograms depict the frequency with which each category was predicted.

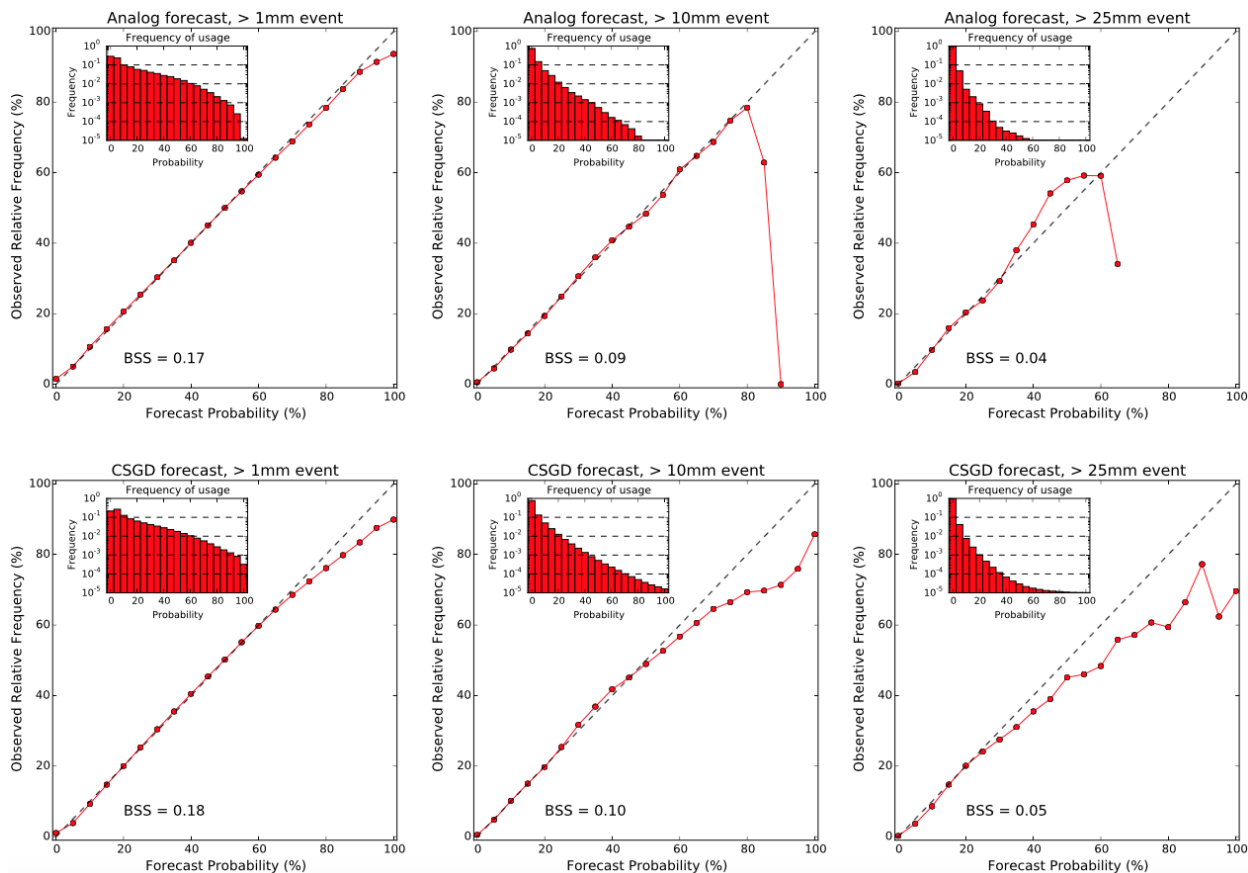


FIG. 11. As Fig. 10 but for +108 to +120 h lead time.

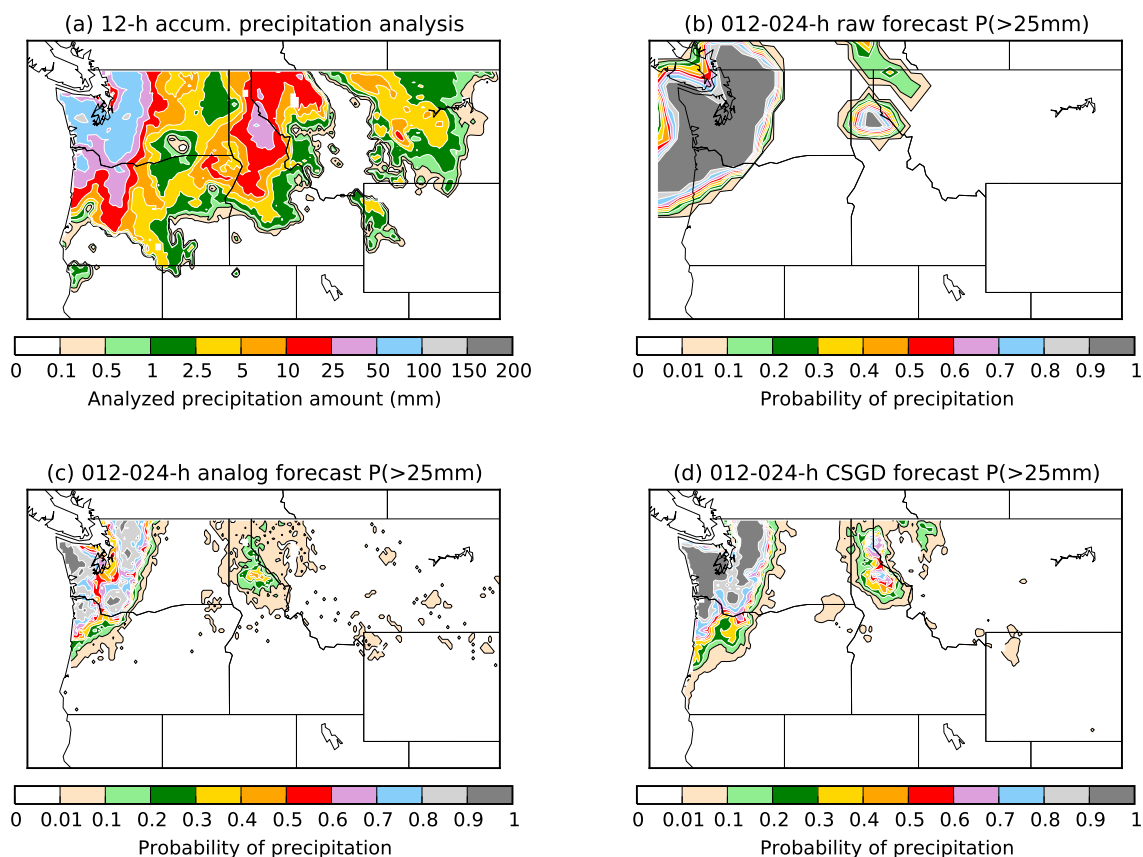


FIG. 12. Analyzed precipitation between 1200 UTC, Nov 6, 2006 and 0000 UTC Nov 7, 2006 (a) and corresponding +12 to +24 h lead probability forecasts for exceeding 25 mm h^{-1} of precipitation by the raw ensemble (b), the analog method (c) and the CSGD regression approach (d).

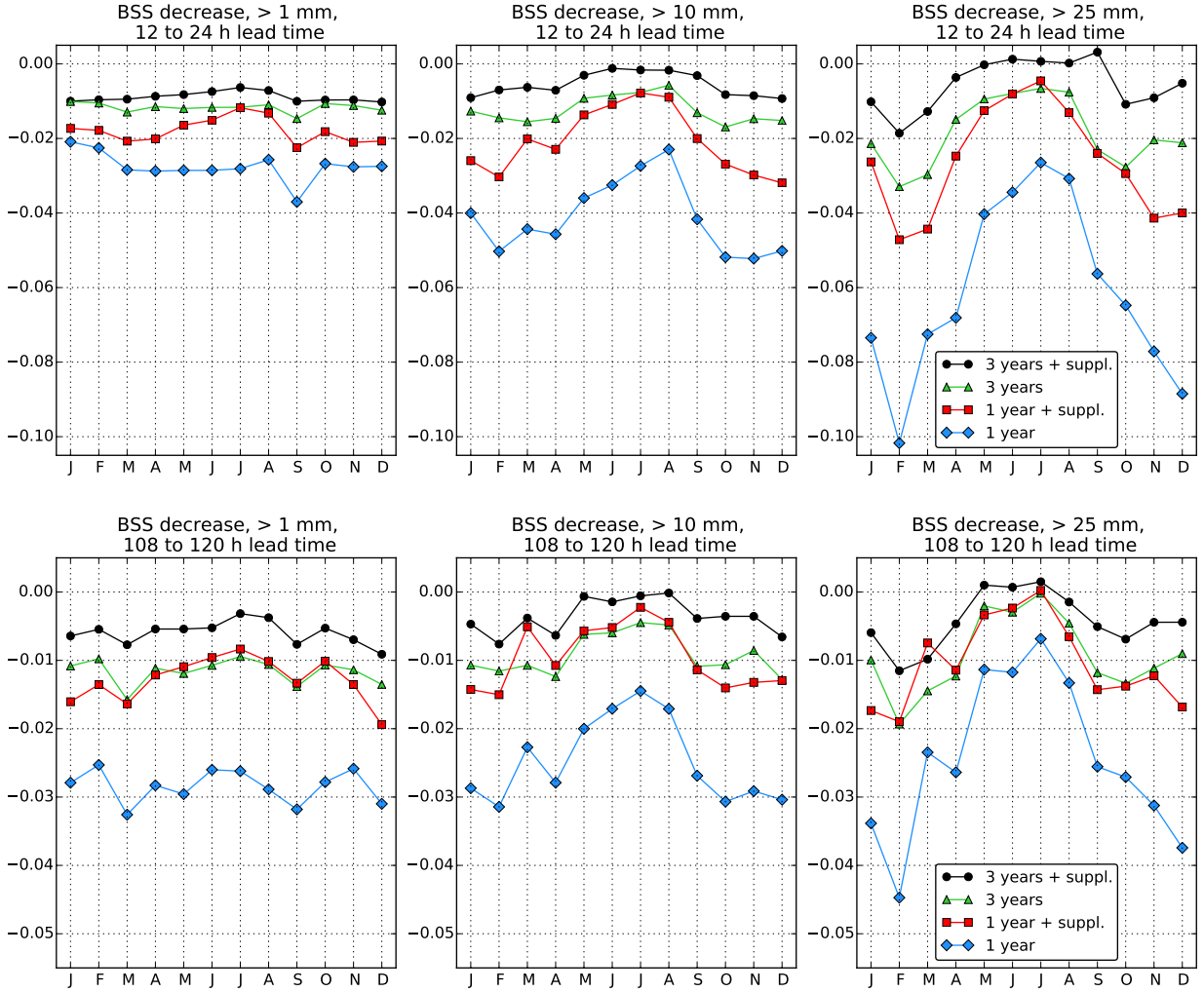


FIG. 13. Decrease of the Brier skill scores (aggregated over all analysis grid points within the CONUS) due to a reduction of the training sample to one or three years of training data. In both cases we also give results for CSGD distributions fitted with additional training data from 19 supplemental locations.

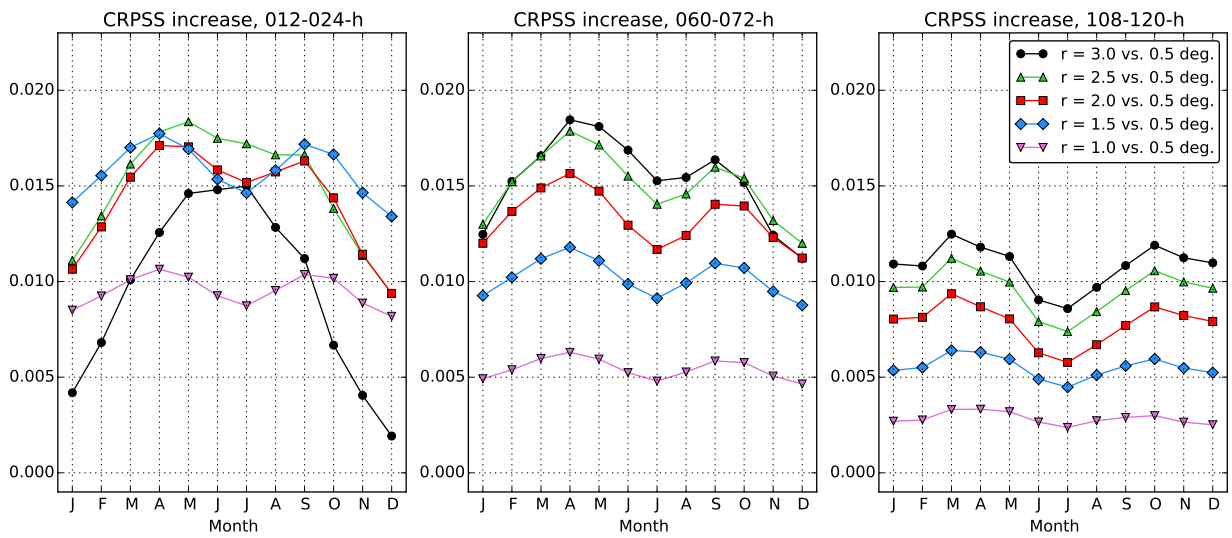


FIG. 14. Increase of CRP skill scores for different neighborhood sizes, relative to the smallest possible neighborhood with $r = 0.5$ degrees.